# **Fetch - Data Analyst Take Home**

**Part 1: Data Exploration**

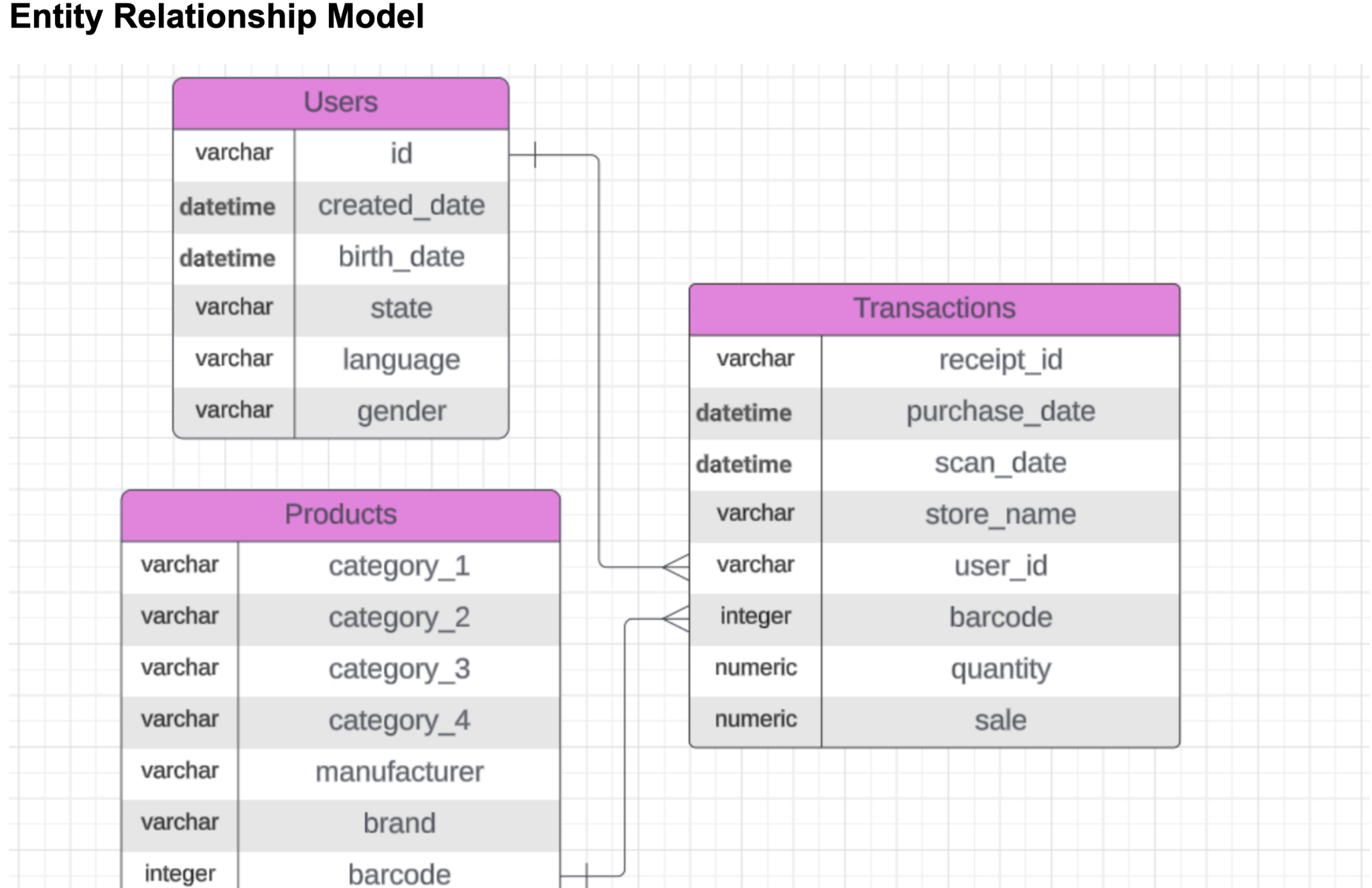
**Question 1 - Are there any data quality issues present?**

Issues with data accuracy, completeness, uniqueness, consistency and dependability are referred to as data quality issues. Inaccurate reporting, poor decision-making, and higher company expenses are all consequences of these problems.

**Introduction to the Data Sets**

We have the Users, Products and the Transaction tables.

Below is a visual representation of the ER Model and how they are connected.

The User table has a unique identifier. The Transaction table and the Products table do not have a primary ID. It was discovered that the receipt\_id as a duplicate value which does not match the requirements of a worthy unique identifier. To be able to move forward with this project. We would consider the barcode on the Transaction table and the Products table as a linking factor. 

I would discuss the data quality issues of these tables below:

1. Data Integrity: Data integrity is the guarantee that information is correct, comprehensive, and consistent for its existence. It entails defending data against corruption, loss, and leakage. We assume that Fetch got the three data sets and it can be trusted.
2. Completeness: The degree to which a dataset includes all relevant information is known as data completeness. It is an indicator of data quality that shows whether any data gaps or missing values exist.

From the USER Data, we could deduce that 'BIRTH\_DATE' has approximately 4% missing values, 'STATE' has approximately 5% missing values, 'LANGUAGE' has about 30.5% missing values and 'GENDER' has approximately 6% missing values.

From the TRANSACTION DATA, All but the 'BARCODE' has no missing values. The 'BARCODE' column with over 11.5% missing values.

From the Product Data, all the columns in the data have missing values. Category\_1 with 0.01%, Category\_2 with 0.17%, Category\_3 with 7.16%, Category\_4 with 92.02%, Manufacturer with 26.78%, Brand with 26.78%, Barcode with 0.48%.

1. Uniqueness: Another metric that can be considered is data quality is data uniqueness, which counts the number of duplicate records in a dataset. It guarantees that every data point is displayed just once.

The user table does not have any data duplicated. Therefore it passed the Uniqueness test.

The Transaction Data could account for a total of 171 duplicate records.

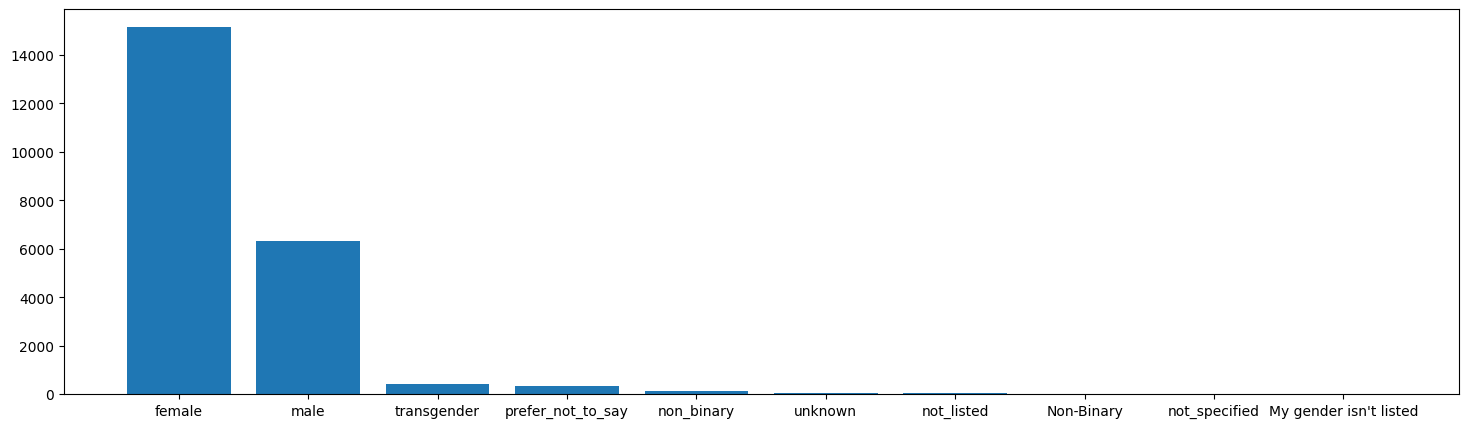
The Product Data have a total of 215 duplicate records from the Data.

**Question 2 - Are there any fields that are challenging to understand?**

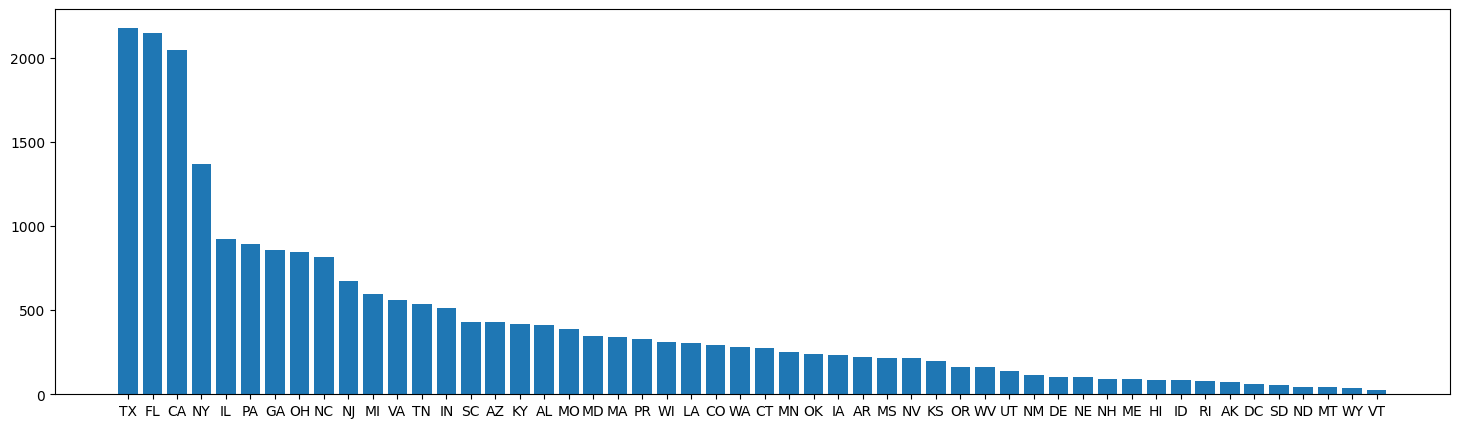
Based on the information provided, the fields are understandable but there are obstacles to shedding light on the dataset, the following are possible interpretations of the fields and the reasons why they might need more explanation:

1. LANGUAGE (Users): Are the languages coded or explicitly named? If code, a key for decoding would be essential.
2. From the product table CATEGORY\_1, CATEGORY\_2, CATEGORY\_3, and CATEGORY\_4 are without descriptions, it's unclear what these categories represent.
3. MANUFACTURER: Does this refer to the total number of manufacturers, or is it a measure of their contribution to the data (e.g., sales or production)? Clarifying the relationship between "MANUFACTURER" and "BRAND" could help, as the terms might overlap conceptually.
4. BRAND: Similar to "MANUFACTURER," is this the count of brands, or does it reflect another metric? If "BRAND" is distinct from "MANUFACTURER," explaining the difference would be useful.
5. BARCODE: "BARCODE" is often associated with unique identifiers for individual products. Does this represent the number of distinct barcodes in the dataset? The purpose and relevance of this field might not be immediately apparent to the audience.
6. BARCODE (Products and Transactions): If this is used as a linking key, its format and uniqueness need confirmation.
7. SCAN\_DATE (Transactions): How is this different from the purchase date? Clarifying this distinction is critical for understanding time-sensitive transactions.
8. FINAL\_SALE (Transactions): Does this include discounts or taxes, or is it the net sale price?

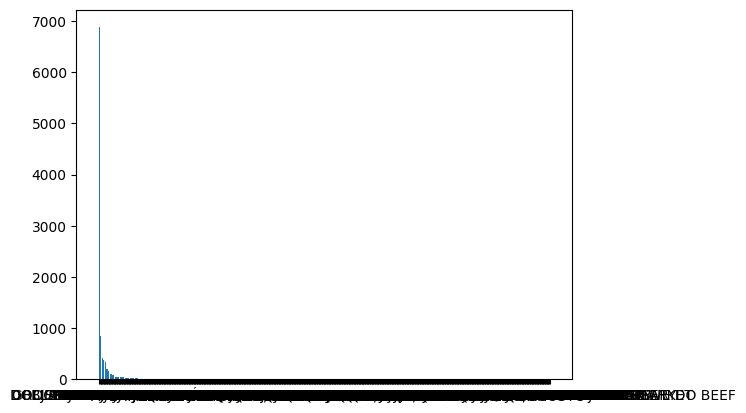
**Data Visualization Examination**



1. Using Barchart to visualize the gender, the Female Gender has the highest value count. The bar chart is used to show the distribution of gender categories based on the number of users in each group. By visually comparing the heights of the bars, you can immediately identify which gender category has the most users.



1. The graphic is dominated by TX, FL, and CA, indicating that these states have the largest counts and values in the dataset. The lowest counts, shown by extremely short bars, are found in states at the far right, such as Vermont (VT), Wyoming (WY), and others. There is a pattern to the distribution, with many states having much lower values and a few states contributing significantly to the total.



1. This is a bar chart showing the STORE\_NAME on the x-axis and their corresponding FINAL\_SALE values on the y-axis. However, the following issues and observations are noticeable. It seems there are too many unique stores, causing overlapping and unreadable labels on the x-axis. Represents the total sales for each store. The chart shows a sharp drop-off after the first few stores, suggesting that a few stores dominate the total sales, while the rest contribute significantly less. The tallest bar represents the store with the highest sales. Subsequent bars are much shorter, indicating a distribution where most stores have low total sales compared to the top performers.

4. By Using Group by it looks at how many transactions each user has made, grouped by their USER\_ID. The key observations are:

1. Top 10 Users: The most active user has made 22 transactions. Several other users have transaction counts ranging between 12 and 20 transactions. These users are your most engaged customers, likely contributing significantly to your overall sales.
2. Bottom 10 Users: Each of these users has only 2 transactions. These users are the least engaged, possibly new customers or those who shop infrequently.

5

This is a bar chart displaying the distribution of product categories from the CATEGORY\_1 field. Here’s an explanation of the visual and its insights: The x-axis represents the unique categories in CATEGORY\_1 (e.g., "Health & Wellness," "Snacks," "Beverages," etc.). Each category is represented as a label.

Frequency: The y-axis represents the number of products that belong to each category. The height of each bar corresponds to the frequency (or count) of products in that category.

Bar Heights: "Health & Wellness" has the tallest bar, indicating it is the most frequent category in the dataset. "Snacks" follows as the second most frequent category. Other categories like "Beverages" and those further down the x-axis have much lower counts, as shown by shorter bars.

**Second Part: SQL Queries**

The SQL Queries were selected from both categories (2 each), with the queries as an attachment in a txt file.

**Closed Ended questions:**

1. What are the top 5 brands by receipt scanned among users 21 and over

**ANSWER:** TRIDENT, CVS, MILK DUDS, SARGENTO, ENTENMANN'S SWEET BAKED GOODS

**Open Ended Questions:**

1. WHO ARE THE POWER USERS OF FETCH

ANSWER: **Assumptions**

1. **Power Users**:
   1. Users with a **high number of transactions** or a **high total spend**.
   2. Thresholds:
      1. Number of Transactions: Top 10% of users by transaction count.
      2. Total Spend: Top 10% of users by total spending.
2. **Data Used**:
   1. TRANSACTION\_TAKEHOME.csv for purchase details.
   2. USER\_TAKEHOME.csv to link transactions to user demographics.
3. **Outcome**:
   1. Identify power users based on their USER\_ID.
   2. Optional: Enrich results with user demographics (STATE, GENDER, etc.).

### **Explanation**

1. **Step 1**:
   * Aggregate data from TRANSACTION\_TAKEHOME to calculate:
     + Total transactions (COUNT(RECEIPT\_ID)) per user.
     + Total spend (SUM(FINAL\_SALE)) per user.
2. **Step 2**:
   * Calculate thresholds for the top 10% of users by transaction count and total spend.
3. **Step 3**:
   * Filter users whose transaction count or total spend meets or exceeds the 90th percentile thresholds.
   * Optionally join with USER\_TAKEHOME for demographic details.

3. AT WHAT PERCENTAGE AS FETCH GROWN OVER THE YEARS

**a. Identify Relevant Data**:

a. Filter transactions to include only the "FETCH" brand. Use the BRAND field from the **Products** dataset and join it with the **Transactions** dataset via BARCODE.

**b. Group by Year**:

1. Extract the year from the PURCHASE\_DATE in the Transactions dataset.
2. Aggregate metrics like FINAL\_SALE or FINAL\_QUANTITY for each year.

C. **Calculate Growth**:

1. Compute year-over-year growth

### **Explanation**

1. **Filter by Brand**:
   * Transactions for "FETCH" are isolated based on the BRAND field in the **Products** dataset.
2. **Yearly Aggregation**:
   * Total sales (SUM(FINAL\_SALE)) are grouped by year using the YEAR function.
3. **Year-over-Year Growth**:
   * The LAG function retrieves the previous year’s sales for each row, enabling growth percentage calculation.
4. **Final Output**:
   * For each year, the query returns:
     + Total sales.
     + Year-over-year growth percentage.

**N.B:**

The SQL queries are saved in a Txt file uploaded along with this document

**Part 3: Communication with Stakeholders**

**Subject:** Detailed Data Analysis Insights and Critical Next Steps

Hi Alex,

I’ve conducted a detailed analysis of the datasets (Users, Transactions, and Products), and I want to share the results, identify challenges, highlight a key trend, and propose steps to address outstanding questions

### **Key Data Challenges and Quality Issues**

1. **Unclear Relationships Between Datasets**:
   * **BARCODE Field**: This field appears in both the Products and Transactions datasets. However, it is unclear if it serves as a fully unique key for linking product information to transactions. Any mismatch here could result in inaccuracies in product-level reporting and trend analysis.
   * **USER\_ID in Transactions**: Several transactions are not linked to user profiles, raising questions about how anonymous or unregistered users are managed. This limits our ability to analyse user-specific behaviours.
2. **Ambiguity in Key Date Fields (Transactions Dataset)**: **SCAN\_DATE vs. PURCHASE\_DATE**: There’s a lack of clarity on the distinction between these fields. For example:
   * Does **SCAN\_DATE** represent the transaction’s recording date, and if so, why might it differ from **PURCHASE\_DATE**?
   * Are there cases where a transaction was scanned much later than the purchase?
   * Understanding this is critical for time-based analysis, such as customer purchasing behaviour or store performance.
3. **Incomplete User Data**:

Many users are missing demographic information such as **BIRTH\_DATE** and **STATE**, which restricts our ability to perform regional or age-based segmentation.

This data gap could skew analyses or leave opportunities untapped in specific segments.

1. **Product Categorization Uncertainty**:

**CATEGORY\_1 to CATEGORY\_4**: These fields in the Products dataset appear hierarchical, but without documentation, it’s unclear whether they represent nested levels of product classification or independent attributes.

This ambiguity hinders category-specific analyses and broader product strategy discussions.

1. **LANGUAGE Field (Users Dataset)**:

If the **LANGUAGE** field uses coded values, we need a mapping to interpret these codes. This impacts our ability to understand language preferences and customize user experiences accordingly.

### **Key Trend Observed**

#### **Sales Dominance of CATEGORY\_4 Products**

* **Over 90% of all sales transactions** are concentrated in **CATEGORY\_4**. This trend raises both opportunities and risks:
  + **Opportunities**: This concentration likely reflects strong alignment with customer preferences, indicating a potential area for strategic reinforcement, such as increased marketing or inventory optimization for this category.
  + **Risks**: Relying heavily on a single category exposes the business to significant risks if consumer preferences change, competitors capture market share, or supply chain disruptions occur. Diversifying product offerings could mitigate these risks and help ensure sustained growth.

### **Outstanding Questions**

1. **Data Definitions and Structure**:

Can you confirm whether **BARCODE** is a reliable key for linking product and transaction data? If not, are there alternative linking mechanisms?

What is the exact distinction between **SCAN\_DATE** and **PURCHASE\_DATE**? Are there specific scenarios where these dates might differ significantly?

Is there a defined hierarchy for **CATEGORY\_1** to **CATEGORY\_4**, and how do these categories map to real-world product groups?

1. **Business Context for Observed Trends**:

Has **CATEGORY\_4** been a strategic focus area recently, perhaps due to promotions or operational decisions?

Are there known external factors (e.g., market trends, seasonal influences) contributing to this concentration in sales?

1. **Addressing Data Gaps**:

Are there existing initiatives to fill in missing user data, especially for the **BIRTH\_DATE** and **STATE** fields? If not, would it be valuable to explore enrichment strategies to improve the dataset?

Should we focus on understanding the sources of missing **USER\_ID** in transactions?

### **Proposed Next Steps**

1. **Clarification and Documentation**: Provide or help create documentation for key fields, including:
   * + **BARCODE** as a linking key.
     + Definitions for **SCAN\_DATE** and **PURCHASE\_DATE**.
     + A description of the product categorization hierarchy.
2. **Trend Validation and Risk Assessment**: Share insights into whether **CATEGORY\_4**’s dominance is intentional or a natural outcome. Explore the potential to reduce reliance on this single category by promoting others.
3. **Data Prioritization**: Decide whether to prioritize resolving missing and ambiguous data or to focus on leveraging observed trends for immediate business strategy.
4. **Collaborative Review**: Schedule a meeting or Slack huddle to discuss these findings, align priorities, and define the next steps for action.

Your input will be invaluable in helping us resolve these open questions and unlock actionable insights. Please let me know if there’s additional context I should consider or if you’d like to schedule time for a deeper discussion.

Thank you for your guidance!

Best regards,  
Bola.